Глава 1

Регрессия — моя профессия

цитата про слонов

хз кто автор

Вы ждали, вы надеялись, вы желали. И вот, наконец, этот момент настал! На подиум выходит она, бесподобная, великолепная, линейная регрессия. Много страниц пришлось прочитать, чтобы наконец то дождаться этого торжественного момента и вкусить её во всей красе.

1.1 А ты точно регрессия?

Как уже ни раз говорилось, байесовские методы это не новые модели, а принципиально другой подход к оцениванию, в рамках которого можно оценить любую старую модель. Надо только лишь выразить своё незнание в виде распределения и найти неплохую выборку. Регрссия не является в этом плане исключением.

Сейчас мы будем иметь дело со слонами. Все формулы будут большими. Простым взглядом со стороны осознать их будет сложновато. Мы призываем читателя взять бумагу, ручку и вывести все результаты вместе с нами.

Начнём. Пусть у нас есть куча наблюдений и целый один регрессор

$$y_i = \beta x_i + u_i, \qquad u_i \sim \mathcal{N}(0, \sigma^2).$$

В такой модели нам необходимо оценить два параметра, коэффициент β и дисперсию σ^2. Предположим, что мы знаем что-то о коэффициенте. Степень нашей уверенности выразим через нормальное распределение. Дисперсия a будет показывать силу нашего незнания, $\beta \mid \sigma^2 \sim N(0, a)$.

Также будем предполагать, что $\sigma \sim InvGamma(s, r)$. Почему распределение будет именно таким? Оно выбирается таким по двум причинам. Первая причина заключается в том, что наши деды придумали сопряжённые распределения. В данном случае обратное гамма-распределение сопряжено с нормальным. Это позволит нам получить приятное апостериорное распределение и сделать задачу его поиска ручной. Безусловно, это очень архаичная причина.

Вторая причина выражать свои априорные мысли именно с помощью этого распределения, более модная и интуитивная. Она вытекает из того как именно моделируют распределение времени, прошедшее между несколькими событиями.

1. Как помнит внимательный читатель, экспоненциальное распределение, $T \sim Exp(\lambda)$, с плотностью распределения $f(t) \propto e^{-\lambda \cdot t}$ обычно используют для моделирования времени между двумя последовательными событиями, которые произошли в Пуассоновском потоке. На самом деле, экспоненциальное распределение это частный случай гамма-распределения, $Exp(\lambda) = \Gamma(0, \lambda)$. Это дат нам почву для обобщений.

2. Гамма-распределение, $T \sim \Gamma(s, r)$, с плотностью $f(t) \propto t^s \cdot e^{-r \cdot t}$ можно использовать, чтобы моделировать время между несколькими событиями, так как такое распределение это ни что иное как сумма $s + 1$ экспоненциального распределения. Есть два способа доказать это. Первый: вспомнить как ищется распределение суммы двух случайных величин. Вспомнив, можно нащупать формулу. Второй: доказать это через характеристические функции. Мы оставим это читателю в качестве упражнения и в традициях лучших книг не будем публиковать решения [1].

3. Обратное гамма-распределение, $T \sim \Gamma^{-1}(s, r)$ с плотностью $f(t) \propto t^{-s} e^{-\frac{r}{t}}$ в таком случае, наверное, описывает частоту серий из несколь-

[1] На самом деле раздражает, когда авторы книг так делают. Многие такие упражнения вообще неочевидно как решаются.

ких событий в течении какого-то времени.

Написать нормально о переходе от времени к частоте и после перейти к дисперсии. Вставить картинок. Сделать нормальных упражнений на гамма-распределение.

Итак, наша текущая задача — найти апостериорное распределение. К несчастью, для сопряжённости, просто выбрать нормальное и обратное гамма распределения недостаточно. Для того, чтобы сделать его совсем-совсем приятным, придётся придумать как именно должны между собой соотноситься a, r и s и ограничить себя тем самым ещё сильнее.

К счастью, STAN позволяет уйти от всех этих ограничений. Мы связываем себе руки только для того, чтобы вручную решить в этой главе задачу по поиску апостериорного распределения, решить парочку упражнений на регуляризаторы, сделать пару интересных наблюдений и ... всё. Дальше мы спокойно скинем с себя оковы. В прекрасном мире будущего все будут свободны, и у каждого будет по два раба.

Формула байеса говорит нам о том, что

$$f(\beta, \sigma \mid y, x) \propto f(y \mid \beta, \sigma^2, x) \cdot f(\beta, \sigma^2 \mid x).$$

Можно сразу же бросаться в бой и прорываться через огромных слонов с экспонентами, как мы это делали в задачке про Машу и Медведей, но мы немного схитрим и возьмём логарифмы от обеих частей равенства

$$\ln f(\beta, \sigma^2 \mid x, y) \propto \ln f(y \mid \beta, \sigma^2, x) + \ln f(\beta, \sigma^2 \mid x).$$

Не забываем держать в голове, что нас интересуют только s, r, β и σ^2. Всем остальным благодаря магии значка $\propto$ мы смело можем пренебрегать. Всё, что будет утеряно, мы восстановим впоследствии из интеграла.

Первое слагаемое — это логарифм нашей функции правдоподобия, второе слагаемое — логарифм нашей априорной плотности распределения. В наших предположениях $f(\beta, \sigma^2 \mid x) = f(\sigma^2 \mid x) \cdot f(\beta \mid \sigma^2, x)$, и второе слагаемое разваливается на два

$$\ln f(y \mid \beta, \sigma^2, x) + \ln f(\beta, \sigma^2 \mid x) = \\ = \sum \ln f(y_i \mid \beta, \sigma^2, x_i) + \ln f(\sigma^2 \mid x) + \ln f(\beta \mid \sigma^2, x).$$

Всё, что мы сделали должно частично избавить нас от слонов и оставить нам только слонят. У нас были такие вот слоны:

$$f(y_i \mid \beta, \sigma^2, x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{1}{2\sigma^2} \cdot (y_i - \beta x_i)^2\right)$$
$$f(\sigma^2 \mid x) = \frac{r^s}{\Gamma(s)} (\sigma^2)^{-s-1} \exp\left(-\frac{r}{\sigma^2}\right)$$

Стали такие вот слонята:

$$\ln f(y_i \mid \beta, \sigma^2, x_i) \propto -\frac{1}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \cdot (y_i - \beta x_i)^2$$
$$\ln f(\sigma^2 \mid x) \propto -(s+1) \ln \sigma^2 - \frac{r}{\sigma^2}$$

С нормальным распределением для β произойдёт точно такая же метаморфоза. Обратите внимание, что из него выскакивает лишнее слагаемое $-\frac{1}{2} \ln a$, на которое нам наплевать. В конечном итоге получаем слонёнка

$$\sum_{i=1}^{n} \left(-\frac{1}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \cdot (y_i - \beta x_i)^2\right) - (s+1) \cdot \ln \sigma^2 - \\ - \frac{r}{\sigma^2} + \underbrace{-\frac{1}{2} \ln a}_{\text{на это нам плевать}} - \frac{(\beta - 0)^2}{2a}.$$

В самом начале мы договорились, что хотим на выходе получить точно такие же апостериорные распределение, нормальное и обратное гамма, но с новыми параметрами, $\mathcal{N}(\tilde{m}, \tilde{a})$ и $\Gamma^{-1}(\tilde{s}, \tilde{r})$.

Новые параметры будут пересчитываться на основании старых и на основании полученных наблюдений.

$$\text{Мы хотим: } -(\tilde{s}+1)\cdot \ln \sigma^2 - \frac{\tilde{r}}{\sigma^2} - \frac{1}{2}\cdot\frac{(\beta-\tilde{m})^2}{\tilde{a}} \tag{1.1}$$

Попробуем сгруппировать нашего гиперслонёнка и получить формулы пересчёта. Соберём все, что находится перед $\ln \sigma^2$ вместе

$$\sum_{i=1}^{n}\left(-\frac{1}{2\sigma^2}\cdot(y_i-\beta x_i)^2\right) - \left(\frac{n}{2}+s+1\right)\cdot\ln\sigma^2 - \frac{r}{\sigma^2} - \frac{\beta^2}{2a}.$$

Ничего другого перед $\ln \sigma^2$ уже не возникнет. Сравним то, что мы получили с тем, что мы хотим и сделаем вывод, что $\tilde{s} = s + \frac{n}{2}$. Первая формула пересчёта обнаружена.

В нашей формуле есть нераскрытый квадрат разности. Как только мы раскроем его, на нас выпрыгнут β и β^2. По аналогии мы можем раскрыть скобки перед квадратом в формуле (1.1), которую мы жаждем получить. Оставшееся уйдёт в константу.

β	$-\frac{1}{2\tilde{a}}$	$-\frac{1}{2a} - \frac{\sum x_i^2}{2\sigma^2}$
β^2	$\frac{\tilde{m}}{\tilde{a}}$	$\frac{1}{\sigma^2}\sum x_i y_i$
Осталось	$-\tilde{r}\cdot\frac{1}{\sigma^2}$	$-\frac{r}{\sigma^2} - \frac{1}{2\sigma^2}\sum y_i^2$

Если просто посмотреть на это один раз со стороны, легко можно не понять откуда взялся результат. Поэтому, ленивый читатель прямо сейчас должен взять ручку и бумагу, вернуться в начало главы и самостоятельно всё вывести, потому что негоже не марать бумагу формулами! А мы, вместе с читателями, которые любят самостоятельно получать монументальные результаты, ещё раз выпишем все формулы пересчёта.

$$\tilde{s} = s + \frac{n}{2} \qquad \tilde{r} = r + \frac{1}{2}\sum y_i \qquad \frac{1}{\tilde{a}} = \frac{1}{a} + \frac{1}{\sigma^2}\sum x_i^2 \qquad \frac{\tilde{m}}{\tilde{a}} = \frac{1}{\sigma^2}\sum x_i y_i$$

Формулы для пересчёта $\tilde{s}$ и $\tilde{r}$ получились хорошими. Остальные две формулы содержат σ^2, которая является гиперпараметром и нам неизвестна. Вспомним о том, что мы хотели наложить на a, r и s какие-то дополнительные ограничения, которые позволили бы нам спокойно осуществлять пересчёт. Судя по всему, нам придётся сделать параметр a пропорциональным гиперпараметру σ^2. Ежеле $a = k \cdot \sigma^2$, то

$$\frac{1}{\tilde{k}\sigma^2} = \frac{1}{k\sigma^2} + \frac{1}{\sigma^2}\sum x_i^2 \quad \Rightarrow \quad \frac{1}{\tilde{k}} = \frac{1}{k} + \sum x_i^2 \quad \Rightarrow \quad \tilde{k} = \frac{k}{\sum x_i^2}$$

$$\frac{\tilde{m}}{\tilde{k}\sigma^2} = \frac{1}{\sigma^2}\sum x_i y_i \quad \Rightarrow \quad \tilde{m} = \tilde{k} \cdot \sum x_i y_i \quad \Rightarrow \quad \tilde{m} = k \cdot \frac{\sum x_i y_i}{\sum x_i^2}.$$

Таким образом, мы получаем удобные формулы для пересчёта параметров апостериорного распределения. Когда наши Деды придумывали это, они понимали, что вывести эти формулы будет нелегко, зато будет легко их использовать. Другими словами, наши Деды восхищались Суворовым.

Обратите внимание на то, что математическое ожидание апостериорного распределения β очень сильно напоминает классическую МНК-оценку этого параметра. Новое значение параметра $\tilde{k}$, в свою очередь напоминает МНК-оценку дисперсии для коэффициента β. Любителям консперологии и теорий различных заговоров на этом моменте пора бы погрузиться в глубокие думы, а также не забыть обвинить Томаса Байеса в масонстве и придумывании плана по захвату мира в 21 веке. Имейте в виду, если вы добросовестно изучите эту книгу, мы найдём вас и предложим пройти обряд по вступлению в наше тайное общество. По толпе собравшихся в нашем тайном месте пронесётся ропот: «Постериор! Постериор!» и на вашем теле появится тайный знак байесовцев.

1.2 А ты точно регрессия в STAN

1.3 Регуляризация для самых маленьких

1.4 Регуляризация и байесовство

1.5 Кросс-валидация без байеса

1.6 Кросс-валидация с байесом

1.7 Ещё раз про приемущества байесовских моделей

1.8 Ещё задачи

Упражнение 1.

Упражнение 2.

Упражнение 3.

Упражнение 4.

Упражнение 5.

Упражнение 6.

1.9 Нет, мама STAN не умеет делать борщ, но он может оценить регрессию

про рег

Тем временем на подиум выходит следующая модель. Раздаются авации. Зал замирает. Красавица, вышедшая на подиум неузнаваема. Каждый чувствует в ней что-то родное, но не может понять кто она... Всему своё время, нужно просто как следует приглядеться.

1.10 О преимуществах байесовских моделей в машинном обучении

Зайдём издалека. Пусть, как всегда, у нас есть выборка (x_i, y_i). Как обычно у нас есть модель

Пусть также $u_i \sim \mathcal{N}(0, \sigma^2 \mid x)$. Это, в свою очередь, означает, что $y_i \sim \mathcal{N}(\beta x_i, \sigma^2)$. Априорно будем считать, что $\beta \sim \mathcal{N}(0, \sigma^2)$.

Пусть в качестве точечной байесовской оценки мы собираемся взять апостериорную моду. Это означает, что мы решаем следующую задачу:

Делаем уже до боли знакомый нам байесовский вывод

$$f(\beta \mid x, y, \sigma^2) \propto f(\beta) \cdot f(y \mid \beta, \sigma^2, x).$$

Сразу же прологорифмируем всё

Уже знакомым нам дв

В тот самый момент, когда она достигла края подиума, толпа всё поняла. У этой красавицы много имён. Кто-то узнал в ней Ridge, гребневая, линейная модель с l_2 регуляризатором — это всё о ней. Авации возобновились.

Наши предположения дали нам метод наименьших квадратов с l_2 регуляризатором. И без того известная нам модель может быть переформулирована на байесовском языке и соответствует достаточно простой вероятностной модели. Этот факт иллюстрирует одно из главных преимуществ байесовского подхода:

Кроме того, байесовские методы позволяют:

1. Строить сложные вероятностные модели из более простых. Байесовский вывод одной модели можно использовать в качестве априорного распределения в следующей вероятностной модели. Так можно скреплять разные модели между собой и строить целые сети.

2. Можно обрабатывать массивы данных, в которых информация поступает последовательно. При поступлении новой порции данных старое апостериорное распределение можно использовать как априорное без необходимости повторного обучения модели с нуля.

3. Возможно использовать априорное распределение, которое предотвращает излишнюю настройку известных параметров на обучающую выборку, это в свою очередь позволяет избежать переобучения.

4. Возможно работать не с полностью размеченными, частично размеченными и с вовсе не размеченными обучающими выборками.

Самое время напрячься и решить парочку упражнений. В части упражнений, связанных с Ridge и Lasso-регрессиями нам снова придётся иметь дело со слонами.

Упражнения

Упражнение 7. Упражнение с пересчётом с циферками

Упражнение 8. Рассмотрим модель

$$y = X\beta + u,$$

где u_i независимы и $\mathcal{N}(0; \sigma^2)$.

Метод гребневой регрессии предполагает минимизацию функции

$$Q(\beta) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{k} \beta_j^2.$$

Рассмотрим байесовский подход к регрессии. Предположим, что априорное распределение имеет вид $\sigma^2 \sim \mathrm{InvGamma}(s, r)$, $\beta_j|\sigma^2 \sim \mathcal{N}(0; a(\sigma^2))$.

При каких s, r и $a(\sigma^2)$ апостериорная мода $\hat{\beta}_{MAP}$ совпадёт с $\hat{\beta}_{Ridge}$?

Упражнение 9. Рассмотрим модель

$$y = X\beta + u,$$

где u_i независимы и $\mathcal{N}(0; \sigma^2)$.

Метод LASSO предполагает минимизацию функции

$$Q(\beta) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{k} |\beta_j|.$$

Рассмотрим байесовский подход к регрессии. Предположим, что априорное распределение имеет вид $\sigma^2 \sim \mathrm{InvGamma}(s, r)$, $\beta_j|\sigma^2 \sim \mathrm{DoubleExp}(a(\sigma^2))$.

При каких s, r и $a(\sigma^2)$ апостериорная мода $\widehat{\beta}_{MAP}$ совпадёт с $\widehat{\beta}_{LASSO}$? При любых s и r, и $a(\sigma^2) = \lambda/2\sigma^2$

Упражнение 10. Упражнения про свойства гамма и экспоненциального распределений с википедии.

1.11 О том как связаны покемоны и регуляризация

Упражнение 11. Храбрый Охотник ловит Покемонов в случайном порядке. Вес i-го пойманного Покемона, y_i, имеет нормальное распределение $\mathcal{N}(\mu; \sigma^2)$. Параметры μ и σ неизвестны.

Храбрый охотник хочет оценить μ по формуле $\hat{\mu} = c \sum_{i=1}^{n} y_i$.

1. При каком c величина $E((\hat{\mu} - \mu)^2)$ будет минимальна?
2. Возможно ли использовать на практике данное c?

$c = \frac{1}{n+\sigma^2/\mu^2}$, нет, так как μ и σ^2 неизвестны.

1.12 Кросс-валидация без байеса

1.13 Кросс-валидация с байесом